Hybrid Latent Semantic Indexing

Iurii Kolomeitsev   Nurlan Shagadatov

Skolkovo Institute of Science and Technology

Machine Learning Project

Document classification

Document-term matrix

matrix of weighted word occurrences in documents (e.g. TF-IDF)

- sparse
- high-dimensional
- low-rank

$\Rightarrow$ dimensionality reduction using Singular Value Decomposition (Latent Semantic Analysis)

- words in different documents share their meaning
- we may know relations between documents

$\Rightarrow$ incorporate additional information in SVD
$\Rightarrow$ we get Hybrid LSI

Problem Statement

Notation

$R \in \mathbb{R}^{D\times T}$ — document-term matrix

$K \in \mathbb{R}^{D\times D}$ — document similarity matrix

$S \in \mathbb{R}^{T\times T}$ — term similarity matrix

Model

Original SVD: $R = U\Sigma V^T$

$$A = RR^T = DCD$$

$$c_{ij} = \cos(i,j) \sim r_i^T r_j \quad \Rightarrow \quad \mathrm{sim}(i,j) \sim r_i^T S r_j$$

$$\begin{cases} RSR^T = U\Sigma^2 U^T \\ R^T K R = V\Sigma^2 V^T \end{cases} \Rightarrow \text{solution (Abdi, 2007)} \quad \widetilde{R} = K^{\frac{1}{2}} R S^{\frac{1}{2}}$$

$\widetilde{U} = K^{\frac{1}{2}} U$, $\widetilde{V} = S^{\frac{1}{2}} V$ — matrices with orthonormal columns

$\Sigma \in \mathbb{R}^{r\times r}$ — diagonal matrix with first r principal values

Computation

Model

$$K^{\frac{1}{2}} R S^{\frac{1}{2}} = \widetilde{U} \Sigma \widetilde{V}^T$$

Efficient Computation

require S, K to be symmetric, positive definite:

$$S = I + \alpha Z, \quad K = I + \beta W,$$

where Z, W — original zero-diagonal similarity matrices with elements satisfying $-1 \leq z_{ij}, w_{ij} \leq 1$

$\Rightarrow$ square root replaced with Cholesky decomposition
$K = L_k L_k^T$, $S = L_s L_s^T$ $\Rightarrow$ final model: $\widetilde{R} = L_K^T R L_S = \widetilde{U} \Sigma \widetilde{V}^T$

Folding-in

r – new document $\quad \Rightarrow \quad u = r L_s \widetilde{V} \Sigma^{-1}$

Local LSI

Local LSI methods integrate the class information and performs separate SVD on the local region of each topic.

Relevant Documents Selecting Method (RDS)

Training algorithm:
For each class c:

1. for documents of class c perform a separate SVD;
2. fold all other documents into the new space;
3. train a classifier of topic c.

- first Local LSI method (Hull, 1994)
- simplest Local LSI method
- suffers from class imbalance

Local Relevancy Weighted LSI (LRW)

Training algorithm:

1. initial classifier of topic c is used to assign initial relevancy score (rs) to each training document;
2. each training document is weighted:

$$f(rs_i) = \frac{1}{1 + e^{-a(rs_i + b)}}$$

$$d_i = d_i * f(rs_i)$$

3. the top $m\gamma$ documents are selected to generate the local term-by-document matrix of the topic c (where m is the number of elements of class c):
4. truncated SVD is performed to generate the local semantic space;
5. all other weighted training documents are folded into the new space
6. all training documents in local LSI vector are used to train a real classifier of topic c.

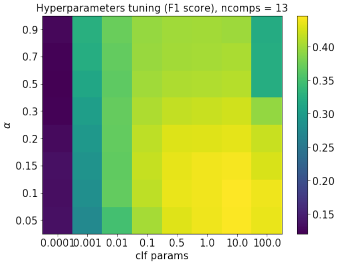
Hyperparameters: a, b, γ

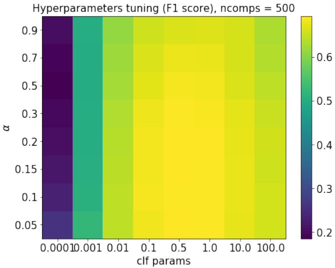
20 Newsgroups

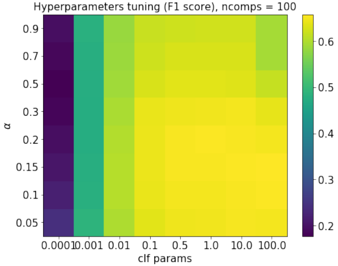
dataset	num docs	avg doc len	initial sparsity, %	sparsity, %
20 Newsgroups	18846	181.6	0.066	0.858

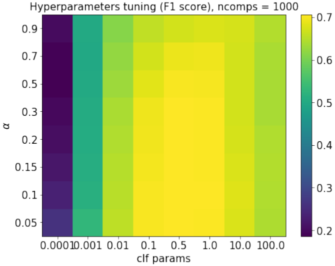
- 20-class classification: news topics
- term similarity: cosine between FastText word representations
- classifier: linear SVM

20 Newsgroups (LSI, HybridLSI hyperparameters choice)

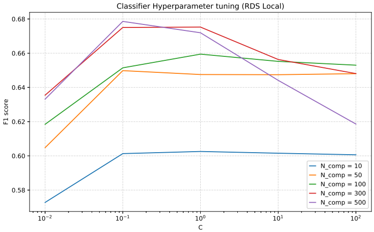


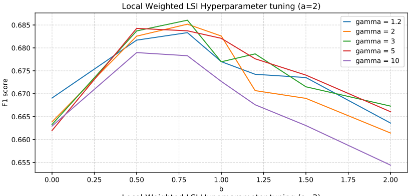


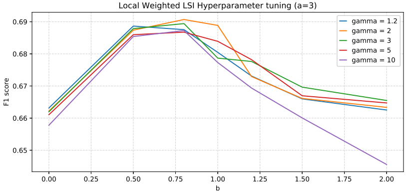


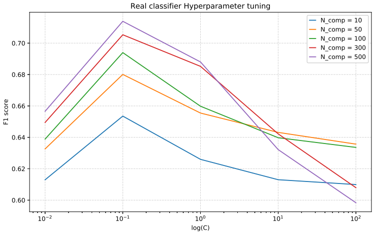


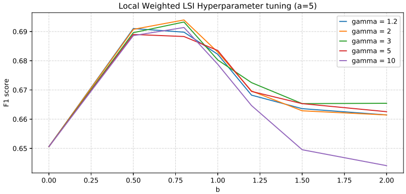
20 Newsgroups (RDS, LRW hyperparameters choice)



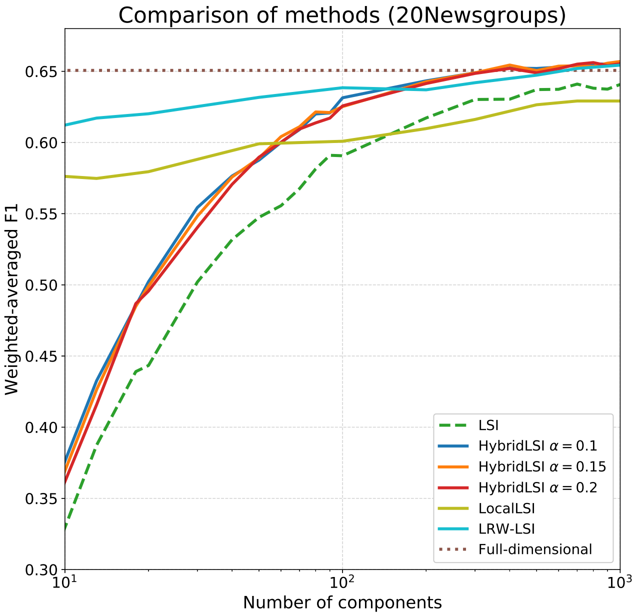








20 Newsgroups

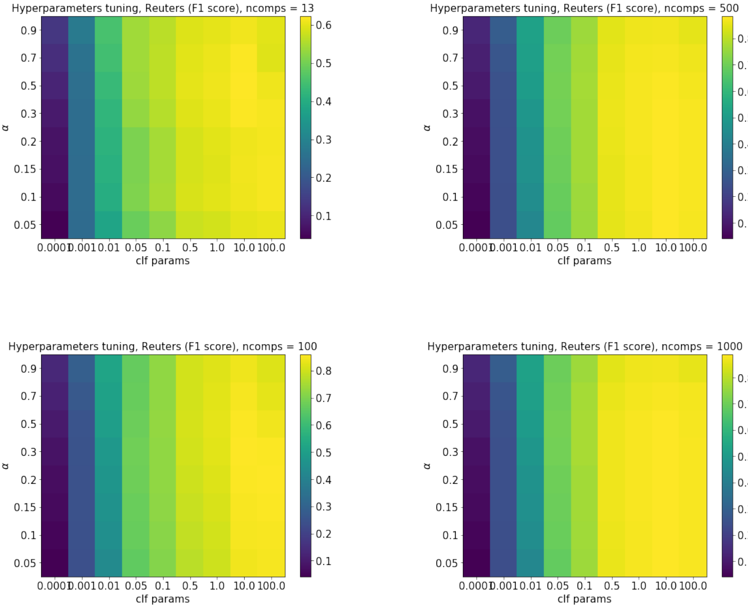


Reuters-21578

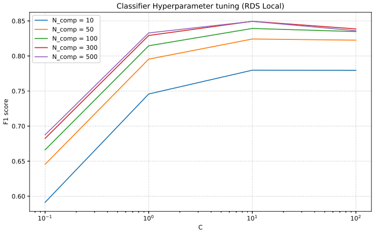
dataset	num docs	avg doc len	initial sparsity, %	sparsity, %
Reuters-21578	10788	127.76	0.195	0.6

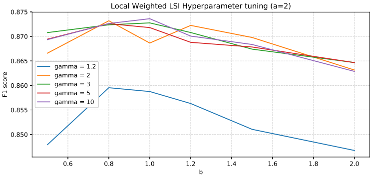
- originally 90-class, multi-label classification: news topics
- considering classes with more than 10 documents, 60 classes remains
- term similarity: cosine between FastText word representations
- classifier: linear SVM

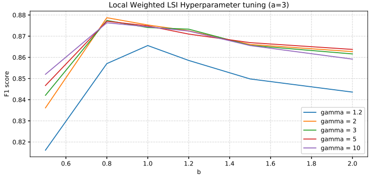
Reuters-21578 (LSI, HybridLSI hyperparameters choice)

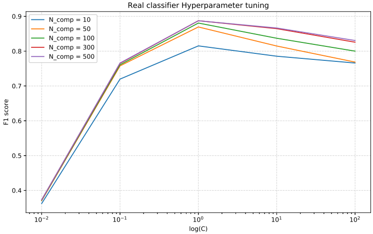


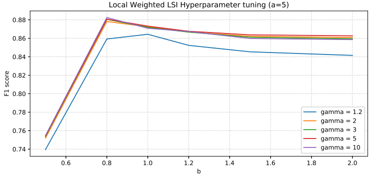
Reuters-21578 (RDS, LRW hyperparameters choice)



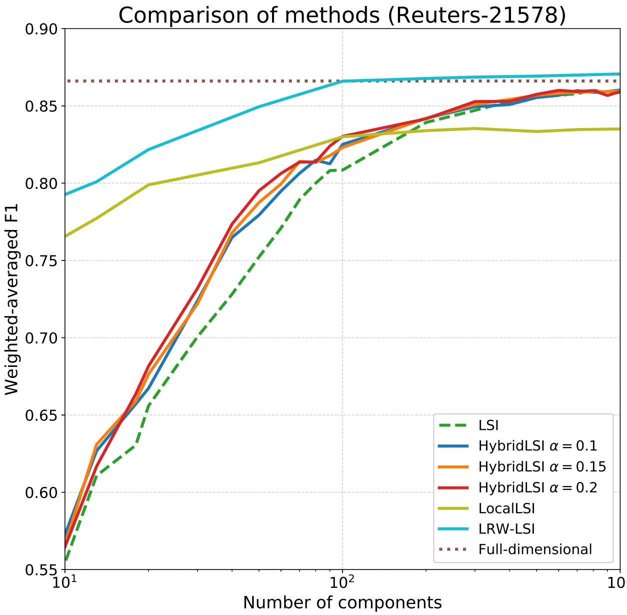




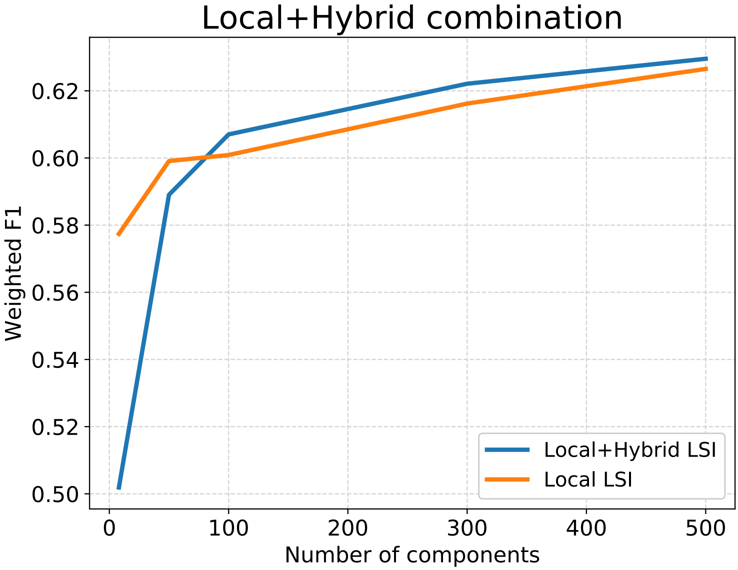




Reuters-21578



Local and Hybrid LSI combination



Summary

- Hybrid LSI model incorporating side information
- different modifications of LSI method have been tested on the 20newsgroups and Reuters-21578 datasets
- Hybrid LSI outperforms standard LSI in all cases
- Hybrid LSI outperforms Local methods in some cases
- Local methods are very good at very low ranks

Future Work

- explore different term similarity measures
- develop approaches to the other text mining problems (e.g. clustering, textual similarity)
- work on the modifications of folding-in
- end-to-end solution where S and K are part of optimization process

Contribution

Iurii Kolomeitsev

- Implementing Hybrid LSI and standard LSI methods;
- testing them on 20newsgroups and Reuters-21578 datasets;
- making report and presentation.

Nurlan Shagadatov

- Implementing RDS LSI and standard LRW LSI methods;
- testing them on 20newsgroups and Reuters-21578 datasets;
- making report and presentation.

References

- E. Frolov and I. Oseledets, HybridSVD: When Collaborative Information is Not Enough, *arXiv:1802.06398*, 2018.
- Herv Abdi, Singular value decomposition (svd) and generalized singular value decomposition (gsvd). *https://www.utd.edu/ herve/Abdi-SVD2007-pretty.pdf*, 2007.
- A. N. Nikolakopoulos, V. Kalantzis and J. D. Garofalakis, EIGENREC: An Efficient and Scalable Latent Factor Family for Top-N Recommendation. *arXiv:1511.06033*, 2015.
- D. Hull, Improving Text Retrieval for the Routing Problem using Latent Semantic Indexing, *Proc. of SIGIR'94*, 1994.